DIST.AR.NET
DISTtributed ARchiving NETwork

Simon Margulies, Ivan Subotic, Lukas Rosenthaler
Imaging & Media Lab
University of Basel
Basel, Switzerland
simon.margulies@unibas.ch http://www.distarnet.ch

The growing production of digital data challenges archiving institutions with new needs for a secure preservation of the cultural heritage of our time. The main subject of the research project 'Distarnet' is to define a protocol for a distributed system for long-term preservation of digital data. The various problems of archiving digital data are analyzed and an implemented solution is presented, taking into account the special needs of archives, museums and libraries being the holders of preservation and distribution of historical source material. Therefore closest attention is paid to the preservation of source material for future scientific researches.

INTRODUCTION

The preservation of digital data is different from the traditional way to preserve data, because digital data itself is meaningless to the naked human eye. Without meaning data is no information. To become understandable for humans, digital data needs to be interpreted and presented by a computer system. Therefore not only the data itself needs to be preserved to guarantee a future readability, but at least also the description for its interpretation by a computer system.

To fulfill these processes archiving institutions need to apply various approaches as outlined in [2]. In summary, a successful solution for the long-term preservation of digital data can only be achieved by a combination of data-carrier migration, data-format migration, emulation and data description. Data-carrier migration and data description are the preconditions for a successful long-term preservation of digital data, because they preserve the data itself and the description needed for its future presentation and interpretation through emulation or data-format migration. Distarnet presents a solution for automated data-carrier migration and supports data description.

OAIS and Distarnet

The OAIS reference model [1] is a widely accepted and used terminology to describe the various processes involved in an archiving institution. The OAIS functional model consist of various entities interacting with each other, as displayed in Figure 1.
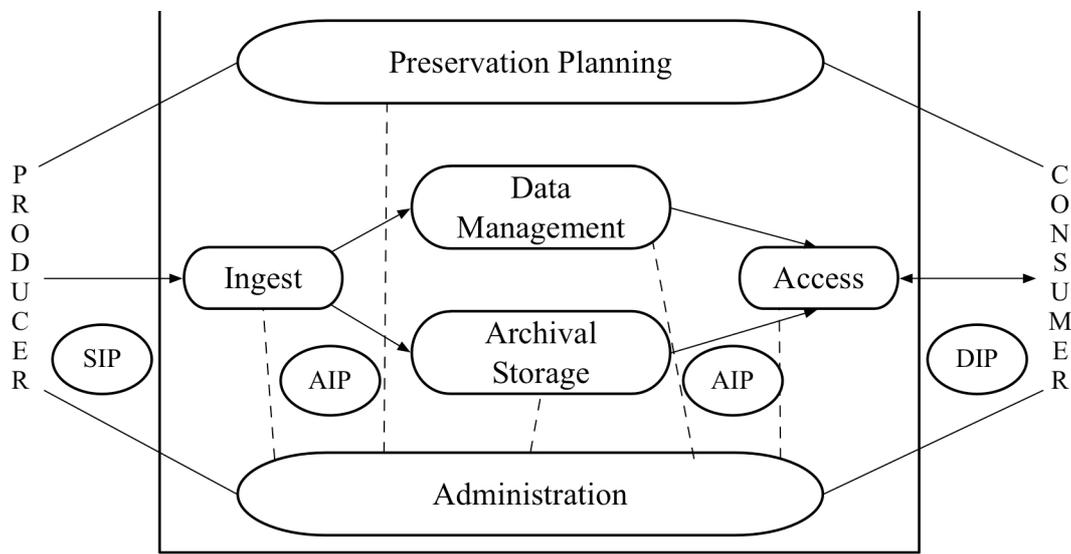
Figure 1: OAIS Functional Entities [1]

The DISTributed ARchival NETwork, Distarnet, is a protocol for a distributed system that offers persistent storage and retrieval for digital data. It corresponds to the OAIS entity Archival Storage, that "provides the services and functions for the storage, maintenance and retrieval of AIPs [Archival Information Package]. Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfill orders."[1] As depicted in Figure 2.
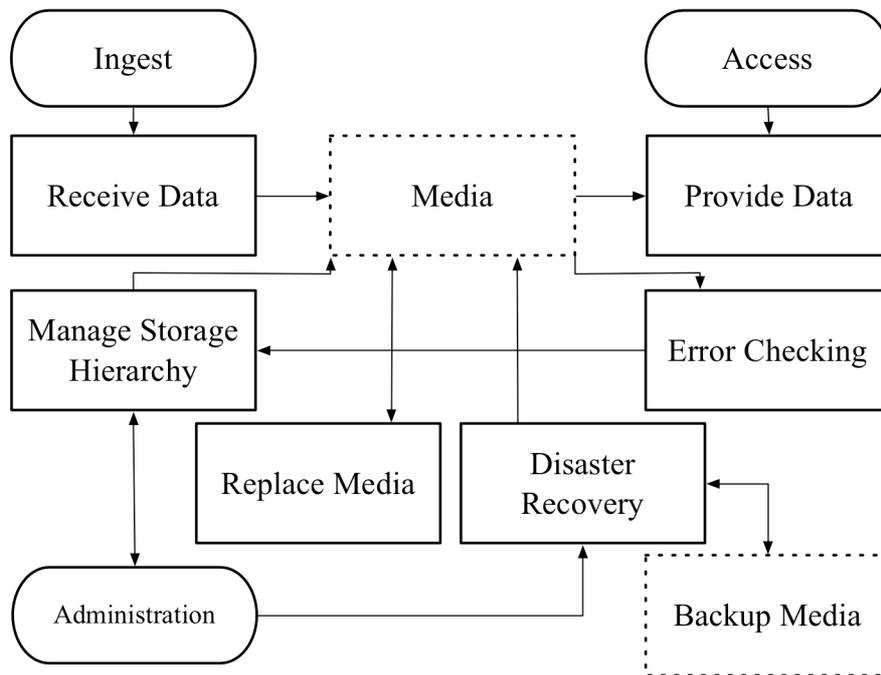

Figure 2: OAIS Functional Entities [1]

To guarantee a high security of the AIPs, the risk of data loss must be minimized and the independence of the AIPs to technological developments must be ensured. Therefore Distarnet supports various levels of data description (Metadata) to make future steps of archiving like emulation or data-format migration, possible. These processes remain external to those of Distarnet.

Being the protocol of a working system Distarnet puts the OAIS definitions of Archival Storage into concrete terms and sets the rules for a distributed system that successfully archives digital data.

THE DISTARNET PROTOCOL 0.2

Distarnet is defined as an XML communication protocol and a set of rules for a distributed system. Its schema will be shared as open source. The system architecture of Distarnet meets the following:

The secure tradition of the data is achieved by building a P2P architecture with strong encryption, controlled redundancy and fault tolerant recovery of network and data. Every node of a Distarnet network communicates in encrypted mode and on top of the TCP/IP protocol. The network stores every AIP in a defined and stable redundancy on different nodes at distant geographical places. If required every node communicates with every other node. The network is fully distributed. All nodes are absolutely equal, so that there is no single point of failure. Status queries to control the availability of stored AIPs are sent periodically between nodes. If a node has lost its data, or if its data appears to be corrupt, the network restores the AIPs by copying them from redundant copies on other nodes. The defined redundancy is reestablished automatically and remains stable. This way not only the secure tradition of the data is assured but the complicated and cost intense data-carrier migration is automated. Carrier-migration becomes almost a non-issue as new hardware can be integrated by simply switching off the old hardware and attaching the new hardware to the network.
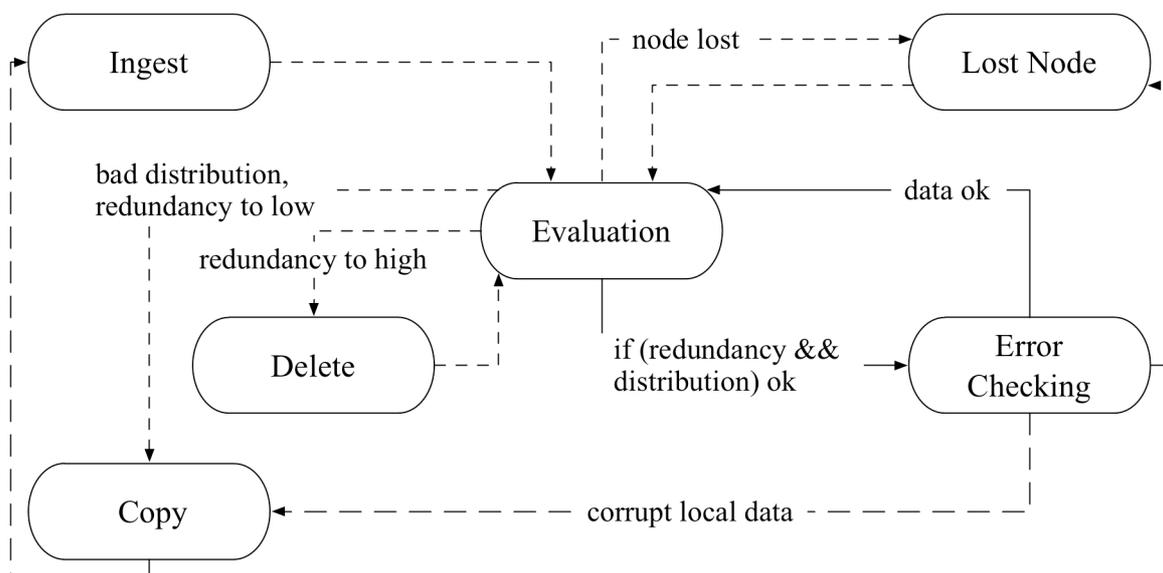

Processes of Distarnet

Figure 3: Processes of a Distarnet-Node

The circular flow of the Distarnet-processes starts upon data ingested. In the next step the AIP is evaluated: according to security criterions, like free space, geographical distribution or node up time, the best node for an AIP is chosen. If the defined redundancy is not met or the distribution of the AIP is not ideal, evaluation starts a copy- or a delete-process. If redundancy and distribution are good the error checking-process starts controlling all local AIPs and their

distant redundant copies upon integrity. If all AIPs are present and unharmed, evaluation starts again. If error checking results negative, either the lost data is recopied to the node itself or the other nodes are informed about the loss of a node (Lost Node). Evaluation starts again to restore the redundancy and to prepare the copy-process. If the redundancy is determined being to high, the delete-process on the concerned node first rechecks, whether this is really the case, before finally deleting the data.

These processes correspond to the OAIS model of Archival Storage as described in table 1.

OAIS	DISTARNET
Ingest / Receive Data	Ingest
Manage Storage Hierarchy	Evaluation
Error Checking	Error Checking
Replace Media	Copy
Disaster Recovery	Copy

Table 1: comparing OAIS-Archival Storage to Distarnet-Processes

From the system behavior of Distarnet and from the fact that digital data is independent of the media it is stored on, it can easily be seen that Media and Backup Media of the OAIS-Archival Storage are a non-issue, respectively replaced by the network itself (and therefore dotted in figure 2). The Provide Data and Access of OAIS are discussed later in this paper under the subsection Metadata.

Security in Distarnet

Distarnet needs security at its lowest level. This means it must not allow communication between nodes or access to data from nodes that are not authorized. To achieve this it uses a Public Key Infrastructure (PKI). PKI is an arrangement, which provides third party vouching for user identities. It also allows the binding of public keys to users. Public keys are used in public key encryption, which is a form of encryption that allows users to encrypt/decrypt a given message without having prior access to a shared secret key. This is done by using a pair of cryptographic keys, which are related mathematically, designated as public and private key. Only the owner knows his private key, his public key is known to all other nodes. A message encrypted with the public key can only be decrypted using the corresponding private key.

For a node to be able to authenticate itself on another node, it will need a public key, which has to be signed by a certification authority, which is mutually accepted by all participants of a certain Distarnet. Such a certification authority (CA) is an entity, which issues the digital certificates for use by all other participant of a Distarnet. It is an example of a trusted third party. This permits the creation of user groups that can choose their own CA, since Distarnet can be used and run by anybody. After the nodes have authorized themselves successfully, a secure connection is established and all traffic between those nodes will be encrypted, which allows for a secure communication over unsecured channels such as the Internet.

Copy-Process in Distarnet

The copy-process in Distarnet is one of its most central processes. It must correspond to the traditional data-carrier migration of digital data. This means that every copy has to be rechecked whether it really has been successful and no data has been lost or written inconsistently during the copy process. For this Distarnet calculates checksum with a function of the Secure Hash Algorithms (currently SHA-1) [4].

Copy in a network means sending files from one node to another. Distarnet does not send the whole file at once, as files could be very big in size: If a copy is started the concerned file is virtually split into a calculated number of chunks depending on the size of the file. A chunk has a network width fixed maximal chunksize (currently 8388608 bytes). There are filesize/chunksize chunks of a file plus the one last chunk only containing the remaining bytes of the file, if there are any. The chunks are numbered from 1 to the calculated number. A node in Distarnet first copies all chunks of a file, then, by putting them together, reconstructs the original file. Before the copy-process the checksum of the whole file and the checksum of every chunk are calculated and send along with the chunks. The receiving node calculates the checksums of the received chunks and compares them with the ones resulted on the sending node. After the collection of all chunks the checksum of the whole file is calculated and compared to the one on the originating node(s). If the check does not result in the same checksum, the chunk or the whole file are requested again.

To speed up the copy-process and to balance the workload for the nodes involved, every node of a copy-process shares already copied chunks with other involved nodes, so that the originating node of a copy-process only needs to copy the file only once into the network.

Metadata in Distarnet

The secure preservation is the precondition of archiving data, but offers neither a guarantee for its readability nor its usability for future scientific interpretation. To fulfill these needs, different types of metadata must be preserved along with their primary data. Through administrative, technical and descriptive metadata, the retrieval, the technical and content-interpretation and consequently readability and scientific usability are made possible. The loss of only one type of metadata can bring along the loss of information about the data and consequently the loss of its readability and usability. In such a case the archiving of the data would have failed.

To face these needs, Distarnet stores data description in RDF (Resource Description Framework) [5] and proposes a basic set of metadata needed to adequately describe the data. Distarnet will offer a mapping of the current standards [like in 6]. It will be possible to add individual schemas and metadata of any participating archiving institution and map them to the schemas already part of Distarnet. This way Distarnet assures the future readability and usability of the data and offers a platform for an overall schema-independent research - corresponding to Provide Data and Access of the OAIS.

Queries in Distarnet

Finding Data and researching its description are crucial to Distarnet, since there is no successful archiving process that securely stores data but cannot provide its findability.

The collection of information in Distarnet is routed over an overlay network that stores information in a distributed hash table (DHT). Distarnet defines a distributed lookup protocol similar to CHORD [3]: Distarnet nodes form a circle by hashing their IP addresses and arranging themselves in an ascending order. Every node keeps track of its direct successor and predecessor node by sending periodical status queries. Additionally every node stores the direct successor of its own successor and the direct predecessor of its own predecessor as backup in case of a lost node and informs them upon a failed status query. Every node manages its part of the DHT and, as backup, stores the part of its predecessor too.

By calculating the hash of the searched information a key is generated and mapped to the DHT - consistent hash functions assure that the calculated keys are distributed regularly and the load of stored information remain balanced among all nodes. The responsible node for that

part of the DHT, the successor of that key, then handles the query and sends back the answer. This responsible node can be found by asking a node of its own shortcut table, CHORDs finger table, that stores some distant nodes, which are responsible for distant hash keys. Finally the responsible node is found by rerouting queries from a distant node to the one actually responsible for the answer. Therewith lookup requires O(log N) messages, with N being the number of nodes participating Distarnet.

THE DISTARNET IMPLEMENTATION

Distarnet is being implemented in Java, currently Version 1.5. The implementation is considered as proof of concept for the protocol. Therefore protocol and implementation are developed simultaneously. The funding of the project will end in December 2007, when protocol version 1.0 will go public.

So far the encrypted communication between nodes, status queries, the ring-topology and the ingest- and copy-process work and are being tested. Next step is to implement the lookup protocol.

References

[1] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002.

[2] S. Margulies, I. Subotic, L. Rosenthaler. Long-term archiving of digital data, DISTributed ARchiving NETwork - DISTARNET. In: EVA 2005 Berlin. Konferenzband, Hg. Gerd Stanke, Andreas Bienert, James Hemsley, Vito Cappellini. Berlin 2005. S. 168-174.

[3] I. Stoica, R. Morris, D. Krager, M. F. Kaashoek, H. Balakrishnan. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications. SIGCOMM 01, August 27-31, 2001, San Diego, California, USA.

[4] http://csrc.nist.gov/CryptoToolkit/tkhash.html

[5] http://www.w3.org/RDF/

[6] http://www.loc.gov/standards/